Automatic Translation Alignment for Ancient Greek and Latin

Tariq Yousef*, Chiara Palladino†, David J. Wright†, Monica Berti*
*University of Leipzig
Augustusplatz 10, 04109 Leipzig, Germany
{tariq.yosef, monica.berti}@uni-leipzig.de

†Furman University
3300 Poinsett Highway, 29613, Greenville SC, USA
chiara.palladino@furman.edu, djwrig85@gmail.com

Abstract

This paper presents the results of automatic translation alignment experiments on text corpus in Ancient Greek translated into Latin. We used a state-of-the-art alignment workflow based on a contextualized multilingual language model that is fine-tuned on the alignment task for Ancient Greek and Latin. The model is fine-tuned on monolingual Ancient Greek texts, bilingual parallel datasets, and manually aligned sentences. The performance of the alignment model is evaluated on an alignment gold standard dataset consisting of 100 parallel fragments aligned manually by two domain experts, with a 90.5% Inter-Annotator-Agreement (IAA). An interactive online interface is provided to enable users to explore the aligned fragments collection and examine the alignment model's output.

Keywords: Translation Alignment, Multilingual Language Models, Evaluation, Alignment Gold Standards

1. Introduction

Translation alignment is the process of finding translation equivalents between a text and its translations. It can be performed at various levels of granularity, from document or paragraph level to word level. It is an important task in Natural Language Processing and Digital Humanities: besides its key role in statistical machine translation (Brown et al., 1993), parallel text alignment has a variety of applications, including cross-lingual annotation projection (Müller, 2017; Xia et al., 2021), language learning (Palladino et al., 2021), and bilingual lexicon induction (Aker et al., 2014; Shi et al., 2021).

Brown et al. (1993) were the first to develop automatic alignment models (*IBM Models*) aiming to extract translation pairs from bilingual corpora. Later, (Och and Ney, 2000) created *Giza++*, an alignment tool based on IBM models and Hidden-Markov alignment models. The continuous efforts made in this field have led to the development of several statistical alignment tools, such as fast_align (Dyer et al., 2013) and EfLoMAl (Östling and Tiedemann, 2016) that outperformed the previous tools on many languages pairs. A new generation of automatic alignment models has emerged with the advances in neural machine translation systems and multilingual contextualized language models. The recent studies employ pre-trained multilingual contextualized word embeddings (Jalili Sabet et al., 2020; Dou and Neubig, 2021) or the attention weights between the encoder and decoder of neural machine translation models (Garg et al., 2019; Chen et al., 2020) to extract translation equivalents from two parallel texts.

1.1. The Challenge of Translation Models for Ancient Languages

In the domain of ancient and generally low-resourced languages, automatic models for translation alignment are still underdeveloped, often due to the lack of large and readily available digitized texts with parallel translations. For Ancient Greek and Latin, the language pair examined in this study, the scarcity is even more staggering, since very little of the hundreds of Latin translations of Greek literature, from the Renaissance to the 19th century, has ever been digitized. Moreover, there are very few manually aligned datasets or gold standards for ancient languages and their translations. These resources are essential to improve automatic translation models, either as training data for automatic methods, or as gold standards against which machine outputs may be tested. To facilitate the collection of alignment pairs and gold standards, various tools have been designed for modern languages (Yousef and Jänicke, 2022). In the case of ancient and low-resourced languages, there are two main web-platforms publicly available: Alpheios [1] and Ugarit [2], which was used in this study [3].

The work presented here uses one of the most extensive digitally available parallel corpora of ancient texts, the *Digital Fragmenta Historicorum Graecorum* (DFHG), which includes over 8000 fragments of Ancient Greek historiographical works and their transla-

[1] https://alpheios.net/.

[2] http://ugarit.ialigner.com/.

[3] The space of this paper does not allow for an extensive description of Ugarit. More information on the tool and its various applications for ancient languages can be found in (Palladino et al., 2021; Yousef et al., 2022)

tions into Latin. We follow the alignment workflow proposed by (Jalili Sabet et al., 2020; Dou and Neubig, 2021), which utilizes contextualized multilingual word embeddings to measure the semantic similarity among the tokens in every two parallel fragments. The contextualized embeddings are generated by a multilingual language model trained and fine-tuned for historical languages. We also created a gold standard dataset annotated manually by two domain experts with alignment guidelines, against which we tested the model's performance. The results are available in an interactive web-based user interface [4] where users can explore the aligned corpus and examine the output of the alignment model. The pre-trained language model is available on `https://huggingface.co/UGARIT/grc-alignment`.

2. The Corpus

The DFHG is the digital open version of the five volumes of the first big printed collection of ancient Greek fragmentary historians edited by Karl Müller in the 19th century[5]. The collection gathers more than eight thousand quotations and text-reuses (*fragments*) of lost works written by more than six hundred authors ranging from the 6th century BC through the 7th century CE (Berti, 2019a; Berti, 2021). Fragments are extracted from still extant source texts and are generally constituted by short passages with information about the relevant lost author and work.

Almost every Greek fragment is translated or shortened into Latin. Limits are of course represented by the fact that the Latin of the corpus is the language used by philologists in the 19th century and not the language of ancient sources. In spite of that, the alignment is very useful not only for translation studies, but also for generating data that can be used for other philological corpora. An example is represented by Named Entities (personal names, places, etc.) that are a strong component of DFHG fragments and that contribute to the creation of authority lists, which are today needed for historical, philological, and linguistic studies (Berti, 2019b). All these characteristics make the DFHG corpus a precious data set for experimenting with translation alignment techniques of ancient languages.

The work described in the following sections has been produced starting with 636 structured XML files of the entire DFHG corpus that are arranged according to volumes and authors of the printed edition and that allow to automatically extract pairs of ancient Greek fragments and their corresponding Latin translations[6].

3. Creating a Gold Standard and Alignment Guidelines

To create the gold standard, 100 fragments randomly selected from the corpus were aligned manually by two experts using Ugarit. An Annotation Style Guide to ensure consistency in the gold standard was also designed in the following way: the two experts, who had previous experience with the alignment of Ancient Greek to Latin in Ugarit, drafted a preliminary set of shared rules together, assessing the most relevant issues (for example, establishing a strategy to manage the presence of articles, which exist in Greek but not in Latin, or defining how to handle enclitics and elliptical constructions). These preliminary rules formed the backbone of the Annotation Style Guide. The experts started the alignment process with a subset of fragments, and discussed issues as they encountered them, revising the Style Guide until it was deemed satisfactory. Then, the experts completed the alignment separately minimizing further discussion, to test the efficiency of the rules defined in the Guide. The gold standard and the guidelines are available on Github[7].

In order to estimate the reliability of the alignment guidelines and the quality of the alignment gold standards, we measured the Inter-Annotator-Agreement (IAA) on the manually annotated fragments, considering the agreement between the annotators on the aligned tokens and the unaligned ones. IAA is a measure that reflects how agreeably multiple annotators can make the same alignment decision for specific tokens.

Ugarit allows annotators to create multi-word alignments (1-to-N, N-to-1, and N-to-N). Therefore, we converted the multi-word alignments to 1-to-1 pairs in order to consider the partial matching of the translation pairs. For instance, the translation pair (A, B C) is considered as two translation pairs (A, B) and (A, C). The resulting IAA is 90.50% and calculated based on equation 1:

$$IAA = 2 * I/(A_1 + A_2) \quad (1)$$

Where A_1 and A_2 be the flattened translation pair sets created by the first and second annotators, respectively, and I is the intersection between them.

To evaluate the performance of automatic alignment systems, (Och and Ney, 2003) proposed two categories of alignments, sure and possible alignments. We followed the same categorization when combining the alignments of the two annotators. We defined sure and possible alignment sets for every sentence as follows:

$$S = A_1 \cap A_2 \quad , \quad P = A_1 \cup A_2$$

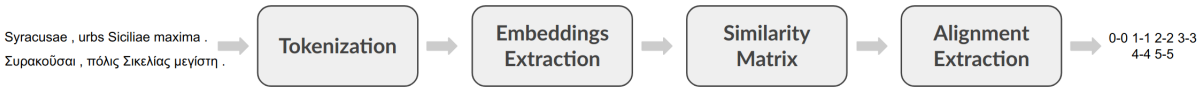
Where A_1 and A_2 are the alignment sets created by the first and second annotators, S denotes sure alignments which include all translation pairs where both annotators agree, P denotes possible alignments where the translation pairs are aligned by at least one annotator.

[4] http://ugarit.ialigner.com/dfhg/

[5] https://www.dfhg-project.org

[6] https://dfhg-project.github.io

[7] `https://github.com/UgaritAlignment/Alignment-Gold-Standards/tree/main/grc-lat`



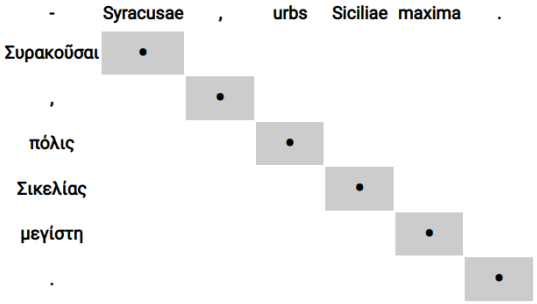
(a) Alignment Workflow

-	_Sy	rac	usa	e	_	,	_ur	bs	_Sicilia	e	_maxima	_	.
_Συ	0.9	0.85	0.85	0.82	0.62	0.62	0.59	0.58	0.62	0.6	0.61	0.6	0.6
ρα	0.84	0.86	0.84	0.81	0.62	0.62	0.59	0.58	0.63	0.62	0.62	0.61	0.61
κο	0.81	0.85	0.86	0.85	0.62	0.61	0.6	0.58	0.63	0.62	0.61	0.6	0.59
ῦ	0.82	0.83	0.87	0.86	0.64	0.64	0.62	0.6	0.63	0.62	0.63	0.6	0.59
σαι	0.81	0.83	0.88	0.89	0.63	0.63	0.62	0.61	0.64	0.62	0.64	0.6	0.6
_	0.58	0.59	0.62	0.63	0.97	0.96	0.67	0.67	0.56	0.59	0.63	0.66	0.64
,	0.58	0.59	0.61	0.63	0.96	0.97	0.67	0.67	0.56	0.59	0.64	0.64	0.63
_	0.58	0.59	0.59	0.59	0.71	0.72	0.93	0.94	0.64	0.64	0.69	0.6	0.59
πόλ	0.58	0.59	0.59	0.59	0.66	0.67	0.94	0.94	0.67	0.66	0.71	0.59	0.58
ις	0.57	0.59	0.6	0.58	0.64	0.66	0.92	0.95	0.67	0.67	0.73	0.6	0.59
_Σι	0.67	0.65	0.66	0.62	0.59	0.58	0.68	0.67	0.92	0.87	0.7	0.61	0.6
κελ	0.63	0.64	0.65	0.63	0.58	0.58	0.67	0.66	0.91	0.9	0.71	0.61	0.6
ίας	0.63	0.66	0.66	0.66	0.6	0.6	0.67	0.66	0.9	0.94	0.72	0.63	0.62
_με	0.61	0.6	0.61	0.6	0.62	0.63	0.7	0.7	0.67	0.67	0.91	0.61	0.59
γί	0.61	0.6	0.61	0.6	0.62	0.63	0.7	0.7	0.66	0.67	0.9	0.59	0.58
στη	0.62	0.62	0.63	0.61	0.62	0.63	0.72	0.73	0.68	0.68	0.92	0.62	0.61
_	0.61	0.61	0.61	0.59	0.66	0.64	0.59	0.59	0.61	0.62	0.63	0.98	0.97
.	0.61	0.61	0.61	0.59	0.66	0.65	0.59	0.59	0.62	0.62	0.63	0.97	0.98

(b) Similarity Matrix (Cosine Similarity)

-	_Sy	rac	usa	e	_	,	_ur	bs	_Sicilia	e	_maxima	_	.
_Συ	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ρα	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
κο	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ῦ	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
σαι	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
_	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
,	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
_	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
πόλ	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
ις	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
_Σι	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
κελ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
ίας	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
_με	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
γί	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
στη	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
_	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

(c) Alignment Extraction (Argmax)



(d) Alignment Result

Figure 1: The alignment process and an example illustrates its workflow.

4. Automatic Alignment

Translation alignment process aims to map word-level equivalents between the source sentence $S = (s_1, s_2, .., s_n)$ and its translation $T = (t_1, t_2, .., t_m)$ (Brown et al., 1993). The process takes S and T as inputs, and produces the set $A = \{(s_i, t_j) : s_i \in S, t_j \in T\}$ where s_i is a translation equivalent of t_j.

Until recently, statistical translation alignment models such as *Giza++*, *fast_align*, and *EfLoMAl* were considered state-of-the-art. However, with the recent advances in language modelling and transformer models, new neural alignment models have been proposed and outperformed the statistical models.

In this paper, we use the state-of-the-art alignment workflow proposed by (Jalili Sabet et al., 2020) and (Dou and Neubig, 2021) which employs pre-trained multilingual contextualized language models to generate word alignments. Further, we fine-tune a language model that can align ancient Greek-English and ancient Greek-Latin with a novel training approach. It combines training over monolingual and bilingual datasets, in addition to supervised training over accurate word-level alignments annotated manually by experts on UGARIT.

4.1. Alignment Workflow

The alignment workflow consists of four main steps (figure 1a): The first step is tokenizing the two parallel sentences into two lists of tokens G and L. Then, extracting embeddings from pre-trained multilingual contextualized language models such as mBERT (Devlin et al., 2018) and XLM-RoBERTa (Conneau et al., 2019) or fine-tuned versions of them for each token. Both models use subword-based tokenization[8], but the tokenization method differs according to the underlying language model. For instance, mBERT uses *WordPiece Tokenizer* whereas XLM-RoBERTa uses a *byte-level BPE* tokenizer. In all experiments, the word embeddings were extracted from the 8th layer of mBERT and XLM-RoBERTa models, since it has achieved the best performance.

The next step is to generate a similarity matrix of size $m * n$ (Figure 1b) where $m = |L|, n = |G|$ and fill it

[8]A tokenization approach splits infrequent words into smaller meaningful subwords. It has shown great performance against word tokenization, especially with multilingual language models, by solving the problems of large vocabulary size and out-of-vocabulary tokens.

using the following formula:

$$\sum_{i}^{n}\sum_{j}^{m} SIM(i,j) = F_{sim}(t^{i}_{grc}, t^{j}_{lat}) \quad (2)$$

Where t^{i}_{grc} is the embedding vector of the ith token in G, t^{j}_{lat} is the embedding vector of the jth token in L, and F_{sim} is a similarity function between the two vectors such as *Cosine Similarity*, *Dot Product*, and *Euclidean distance*.

Once the similarity matrix is computed, alignments can be extracted by applying an extraction algorithm (Figure 1c). (Dou and Neubig, 2021) proposed two probability thresholding-based methods to extract alignments from the similarity matrix, namely, $Softmax$ and $Entmax$ (Peters et al., 2019). Dou and Neubig (2021) applies the extraction in two directions and then considers the intersection between them. Moreover, (Jalili Sabet et al., 2020) proposed three methods including $Argmax$, a baseline method , $Itermax$, an iterative method, and $Match$, a graph-based method.

The last step of the alignment workflow is to convert subword-level alignments to word-level alignments. For this purpose we follow the heuristic principle *"two words are aligned if any of their subwords are aligned"* as in Jalili-Sabet et al. (2020), (Zenkel et al., 2020), and Dou and Neubig (2021) (Figure 1d).

4.2. Language Models

The existing multilingual contextualized language models mBERT and XLM-RoBERTa are not trained on ancient Greek texts but on modern Greek, which is very different. Therefore, we had to train and fine-tune them with ancient Greek texts to enable them to process ancient Greek texts. To this end, we propose a training approach that consists of three main phases:

- ***Ex1***: in this initial phase, we train the models on 12 million Ancient Greek tokens with Masked Language Model (MLM) training objective. The training dataset is extracted from the Perseus Digital Library, the First1KGreek Project[9], and the PROIEL, PERSEUS[10], and Gorman [11] treebanking datasets.

- ***Ex2***: in this phase, we perform unsupervised fine-tuning of models obtained from the previous phase using 32500 Ancient Greek-English parallel sentences taken from the Perseus Digital Library[12] (*Iliad*, *Odyssey*, *Xenophon*, *New Testament*), in addition to 8000 Ancient Greek-Latin parallel fragments (DFHG Corpus)[13], with 4000 further parallel sentences taken from UGARIT database. The texts are in different languages, mainly Ancient Greek-English, Ancient Greek-Latin, and Ancient Greek-Georgian. The training objectives used in this phase are: Masked Language Model (MLM), Translation Language Modeling (TLM), Self-training Objective (SO), and Parallel Sentence Identification (PSI).

- ***Ex3***: in this phase, we perform supervised training with Self-training Objective (SO) to the fine-tuned models obtained after *EX2* using manually word-level aligned dataset provided by UGARIT. The alignments are accurate and clean since they are done by scholars, teachers, and experts. The dataset consists of 2265 parallel texts and almost 100k translation pairs.

The training objectives used in the experiments are proposed by (Dou and Neubig, 2021).

4.3. Evaluation

We evaluated the performance of the proposed alignment workflow based on our fine-tuned language models against the alignment gold standard by employing $Precision$, $Recall$, $F1$, and Alignment Error Rate (AER) which can be computed as in equations 3.

$$Precision = \frac{|A \cap P|}{|A|} \quad , \quad Recall = \frac{|A \cap S|}{|S|}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

$$AER = 1 - \frac{|A \cap P| + |A \cap S|}{|A| + |S|}$$

Where A indicates the alignments set predicted by the model, P and S indicate respectively the $Possible$ and $Sure$ alignment sets in the gold standards, and $|.|$ denotes the length of the set.

As baseline models, we used *Giza++*, *fast_align*, and *EfLoMAl* with their default parameters trained on the whole DFHG dataset.

Table 1 shows poor performance for the statistical models *Giza++* and *fast_algin* since they require a vast parallel corpus, and because of the high number of unique word forms in the corpus (66% of the ancient Greek words and 59% of the Latin, Table **??**).

Further, The table shows that the baseline models outperform the zero-shot XLM-RoBERTa and mBERT with all extraction algorithms, which is understandable since both models are trained on modern Greek, which differs significantly from ancient Greek. The results also show that training the models on monolingual ancient Greek texts (Ex1) enhanced the performance of the alignment workflow and reduced the *AER* significantly. Both models at this point outperformed *Giza++* and *fast_align* but underperformed *EfLoMAl*.

Further performance enhancement is accomplished by fine-tuning the models with bilingual sentences (Ex2); the model outperforms all baseline models significantly. Moreover, the remarkable enhancement has been achieved by incorporating supervised signals by fine-tuning the models on word-level manually aligned parallel texts (Ex3) with the Self-training Objective (SO). SO encourages the aligned words to have closer

[9] https://opengreekandlatin.github.io/First1KGreek/

[10] https://universaldependencies.org

[11] https://vgorman1.github.io/

[12] https://github.com/PerseusDL/canonical-greekLit

[13] The 100 fragments used as gold standard are excluded.

		Precision	Recall	F1	AER
Baseline	Giza++	55.03%	67.61%	60.67%	39.48%
	fast_align	51.64%	70.51%	59.62%	40.67%
	EfLoMAl	76.79%	78.12%	77.45%	22.57%

		XLM-RoBERTa				mBERT			
		Precision	Recall	F1	AER	Precision	Recall	F1	AER
Zero-Shot	Softmax	49.35%	42.10%	45.44%	54.49%	55.40%	51.52%	53.39%	46.55%
	Argmax	62.10%	41.88%	50.02%	49.77%	80.25%	34.86%	48.61%	50.87%
Ex1	Softmax	63.79%	57.61%	60.54%	39.40%	65.89%	69.49%	67.64%	32.41%
	Argmax	75.15%	59.20%	66.23%	33.61%	81.20%	55.43%	65.88%	33.84%
Ex2	Softmax	80.89%	82.68%	81.78%	18.24%	82.48%	83.91%	83.19%	16.83%
	Argmax	86.71%	81.74%	84.15%	15.79%	87.94%	78.55%	82.98%	16.90%
Ex3	Softmax	88.94%	89.13%	89.03%	10.97%	85.67%	84.64%	85.15%	14.83%
	Argmax	91.49%	87.32%	89.36%	**10.60%**	90.15%	78.26%	83.79%	16.09%

Table 1: Evaluation Results, The evaluation was conducted using the five extraction approaches, but we mentioned only the top two.

contextualized representations, increasing their semantic similarity. We also noticed that supervised training had a greater impact on the performance of fine-tuned XLM-RoBERTa than fine-tuned mBERT model.

Figure 2 shows a visual evaluation (Yousef and Jänicke, 2022) of the output of two alignment extraction approaches based on the fine-tuned XML-RoBERTa language model of Ex3. The agreement is shown in green color, big and small dots denotes gold standards sure and possible alignments. As we can see, $Softmax$ predicts more translation pairs than $Argmax$, and $Argmax$ output is a subset of $Softmax$ output, which explains why $Softmax$ outperforms $Argmax$ regarding the Recall and underperforms it regarding the Precision. A full comparison of different alignment models over the gold standard dataset is available under `http://vis4nlp.com/alignmenteval/`.

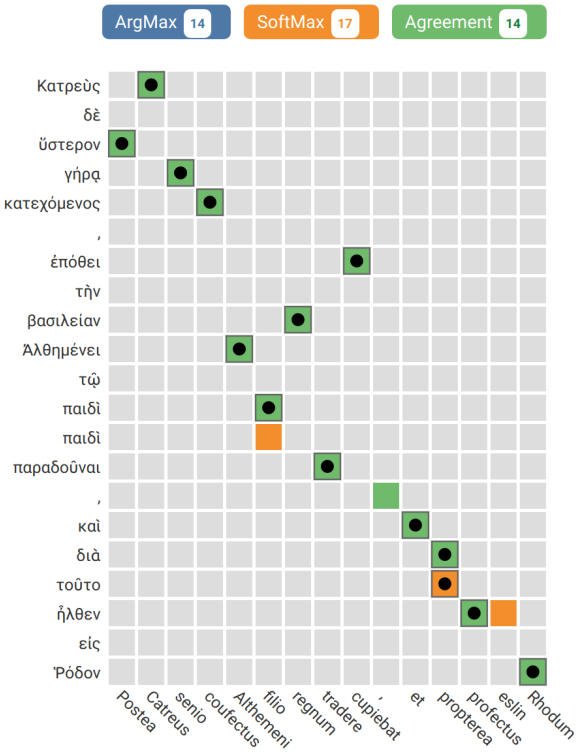


Figure 2: Alignment model (Ex3) output with two alignment extraction approaches compared to the gold standard.

4.4. Qualitative Evaluation

While quantitative evaluation provides a summarized overview of the quality of the models, it fails to provide an in-depth analysis of performance limitations, strengths, or frequent alignment errors. Therefore, we conducted a qualitative evaluation of the alignment output on 50 random fragments, performed by a domain expert.

The evaluation subset includes a total of 748 translation pairs with 40 incorrect pairs (5.35%). The model correctly aligned 54 of 54 prepositions (100%), 18 of 18 adverbs (100%), 186 of 188 Named-Entities (98.94%), 53 of 54 adjectives (98.15%), 53 of 54 conjunctions (98.15%), 40 of 41 pronouns (97.56%), 119 of 125 verbs (95.20%) and 125 of 133 substantives (93.98%).

Most recurrent errors are due to the absence of articles in Latin: Greek articles are sometimes incorrectly aligned with contextual Latin adjectives, pronouns, and substantives. Other limits are also due to elliptical constructions, where finding a certain match is more complex. Finally, Greek particles are variously aligned with Latin conjunctions and adverbs.

5. Conclusion

In this paper, we fine-tuned a multilingual language model that can align ancient Greek and Latin texts following a state-of-the-art alignment workflow. Moreover, we created a gold standard dataset to evaluate the model's performance. Both quantitative and qualitative evaluations confirmed the good performance of the model.

The main challenge we encountered was aligning long fragments. Since most of the fragments are long (over 100 tokens/fragments), there is a need to develop better text segmentation or sentence level alignment models. Further, this study was limited to the specific dataset of the DFHG, which is one of the largest digitized GRC-LAT parallel corpora available. However, in the future, we plan to include more diverse datasets, e.g. expanding towards other literary genres, such as poetry, by scouting available digital libraries and implementing our collaboration with Ugarit users who work on the alignment of these two languages. In addition, we also plan to expand the model and train it to include more language pairs such as ancient Greek-Italian, ancient Greek-French and further.

Acknowledgment

This project is developed thanks to the important contribution of our community of scholars and language learners: Gregory R. Crane, Chiara Palladino, Monica Berti, Farnoosh Shamsian, Maia Shukhoshvili, Anise D'Orange Ferreira, David J. Wright, Christopher Blackwell, Clifford Robinson, Brian Clark.

6. Bibliography

Aker, A., Paramita, M. L., Pinnis, M., and Gaizauskas, R. (2014). Bilingual dictionaries for all eu languages. In *LREC 2014 Proceedings*, pages 2839–2845. European Language Resources Association.

Berti, M. (2019a). Historical fragmentary texts in the digital age. In Monica Berti, editor, *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, pages 257–276. De Gruyter, Berlin.

Berti, M. (2019b). Named entity annotation for ancient greek with inception. In Kiril Simov et al., editors, *Proceedings of CLARIN Annual Conference 2019*, pages 1–4, Leipzig, Germany. CLARIN.

Berti, M. (2021). *Digital Editions of Historical Fragmentary Texts*. Digital Classics Books 5. Propylaeum, Heidelberg.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chen, Y., Liu, Y., Chen, G., Jiang, X., and Liu, Q. (2020). Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, November. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Dou, Z.-Y. and Neubig, G. (2021). Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online, April. Association for Computational Linguistics.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June. Association for Computational Linguistics.

Garg, S., Peitz, S., Nallasamy, U., and Paulik, M. (2019). Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4462, Hong Kong, China, November. Association for Computational Linguistics.

Jalili Sabet, M., Dufter, P., Yvon, F., and Schütze, H. (2020). SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online, November. Association for Computational Linguistics.

Müller, M. (2017). Treatment of markup in statistical machine translation. Association of Computational Linguistics.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, page 440–447, USA. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Östling, R. and Tiedemann, J. (2016). Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146, October.

Palladino, C., Foradi, M., and Yousef, T. (2021). Translation alignment for historical language learning: a case study. *Digital Humanities Quarterly*, 15(3).

Peters, B., Niculae, V., and Martins, A. F. (2019). Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519.

Shi, H., Zettlemoyer, L., and Wang, S. I. (2021). Bilingual lexicon induction via unsupervised bitext construction and word alignment. *CoRR*, abs/2101.00148.

Xia, P., Qin, G., Vashishtha, S., Chen, Y., Chen, T., May, C., Harman, C., Rawlins, K., White, A. S., and Van Durme, B. (2021). Lome: Large ontology multilingual extraction. *arXiv preprint arXiv:2101.12175*.

Yousef, T. and Jänicke, S. (2022). Visual evaluation of translation alignment data. In *Proc. EuroVis*, volume 22.

Yousef, T., Palladino, C., Shamsian, F., and Foradi, M. (2022). Translation alignment with ugarit. *Information*, 13(2).

Zenkel, T., Wuebker, J., and DeNero, J. (2020). End-to-end neural word alignment outperforms GIZA++. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1605–1617, Online, July. Association for Computational Linguistics.